微博用户标签与博文内容相关度研究*

朱 玲 薛春香 章成志 傅 柱
(南京理工大学经济管理学院 南京 210094)

摘要:【目的】探索微博用户标签与其发布微博主题之间的潜在关系,为微博类应用平台的主题发现以及用户标签自动推荐服务提供参考。【方法】利用爬虫程序抓取"自然语言处理"领域新浪微博用户信息及微博,对抓取的微博内容进行分词并对用户标签进行语义扩充,运用编辑距离算法将标签集与用户的微博内容进行匹配。【结果】对匹配结果进行抽样分析,发现新浪微博平台上,学术领域微博用户标签和用户所发微博内容具有一定的相关度。【局限】仅对学术领域和新浪微博进行相关研究,研究领域和应用平台有待进一步扩展。【结论】微博标签推荐系统可以将用户微博内容作为标签推荐的重要数据来源,为用户提供更有针对性的个性化标签;同时,在对微博内容进行主题抽取和分析时,可以借助微博用户标签优化分析结果。

关键词:微博主题分析 用户标签 相关度度量 主题标引 用户建模
分类号:G203

1 引 言

目前,社会化标签系统已成为互联网最流行的在线服务之一。社会化标签在社区发现、信息推荐、信息集成等方面均具有重要价值。微博用户标签是用户依据其所在领域或个性爱好给自己做的标记,在体现个性化特征的同时也给微博好友推荐、用户社区细分等提供丰富的信息来源。微博用户标签与用户所发微博有一定关联,考察微博用户标签与微博内容的关联程度,对微博用户标签自动推荐、好友推荐、博文标签推荐、博文主题检索以及信息推荐服务具有重要意义。

然而,目前用户标签与微博内容之间相关程度的量化研究不多。基于此,本文以新浪微博中的学术型用户为例,采集用户的标签和微博内容,利用这些数据进行用户标签与微博内容关联度的统计分析。该研究可进一步丰富信息组织领域的研究内容,并为微博类应用平台的用户标签自动推荐服务提供参考,以提高微博应用的服务质量。

2 研究综述

早期,标签与文献相关度的研究主要集中在网页、图书和期刊论文等较为正式的信息资源上,通过比较标签与文献主题词或关键词的相似度来衡量。2006年,Al-Khalifa等[1]利用Yahoo关键词抽取工具抽取网页关键词,并将机器抽取的关键词集合、大众标注的标签集合、专业标引人员的标注结果三者进行两两匹配,结果表明,专业标引人员的标注结果与大众标注的标签的重合度要高于机器抽取的关键词。2008年,通过对比主题词和社会化标签,Rolla[2]发现大众标注对图书的描述更加全面细致,能够提高书目检索性能,而主题词只能作为图书基本信息的标引。2009年,Thomas等[3]的相关研究也得出了相同的结论。2010年,Lu等[4]将LibraryThing上用户对图书标注的标签与其在图书馆中标注的Library of Congress Subject Headings主题词进行比较,发现用户标注的标签可以提高图书馆资源的可获取程度。同年,潘婵等[5]以

通讯作者:薛春香,ORCID:0000-0002-5729-819X,E-mail:xuechunxiang@njust.edu.cn。
*本文系国家自然科学基金项目"基于聚合的社会化短文本信息处理与细粒度倾向性分析"(项目编号:71503126)、国家社会科学基金项目"在线社交网络中基于用户的知识组织模式研究"(项目编号:14BTQ033)和江苏省社会科学基金项目"新媒体环境下报纸新闻信息资源开发利用研究"(项目编号:14TQB10)的研究成果之一。

Del.icio.us 为平台，分析了标签和关键词之间的差别，发现娱乐领域和学术领域的标签与关键词的相似度有很大差别。但是他们的研究范围主要针对娱乐和学术两大领域，选取调查的数量也不多，每类只选取了十几个对象。2011 年，Kipp[6]以学术期刊上的文章为数据来源，从用户标签、作者关键词和主题词这三方面收集学术期刊上的文章，通过描述性统计等措施发现关键词和用户标签的匹配存在差异。2012 年，Lee 等[7]将 Medline 数据库中 231 388 篇论文的 MeSH 主题词与 CiteULike 上用户赋予其的标签进行比较，认为社会化标注不能代替传统的受控标引。

随着微博的出现，一些学者开始以其为对象，对微博内容和标签的相关性进行研究。黄红霞等[8]以微博为研究对象，通过对比用户标签与机器标签，发现用户的微博内容与其用户标签有一定关联。章成志等[9]以腾讯微博为研究对象，调研用户标签的主题表达能力，结果表明该平台上有用户标签且影响力较高的用户，约 1/3 的用户标签与微博内容关键词有关。邢千里等[10]假设标签和微博都能够表示用户所关注的主题，对用户标签内容与微博内容之间的关系进行研究，发现标签越相似的用户，微博内容也越相似。

综上，国内外学者已经对文本主题和标签进行了初步研究，指出在特定领域中文本主题和标签既相关又存在一定差异。目前此类研究主要集中在学术和娱乐这两大领域中。此外，对微博用户标签与微博内容关联程度定量方面的研究，均是将用户所发微博看作一个集合来分析，缺少以单条微博为对象的研究。

3　研究设计

3.1　基本思路

本文利用爬虫程序抓取“自然语言处理”领域新浪微博用户信息及微博，通过用户标签与用户微博博文的关联分析以探测微博用户标签与其发布微博的主题相关性，从而为基于用户标签的微博主题识别提供可能性。研究思路如图 1 所示。

(1) 微博用户数据采集。选择新浪微博“自然语言处理”领域的用户为研究对象，以“自然语言处理”、“中文信息处理”为关键词，通过新浪微博 API 接口抓取该领域 835 人的用户信息以及 735 359 条微博数据，其中用户信息涉及用户 ID、昵称、性别、用户标签等。

(2) 用户标签扩充。为了更好地表示出与标签相关领域的内容，提高语义匹配效率，利用清华大学智能技术与系统国家重点实验室信息检索组梁斌博士研发的词库 API①对用户的每个标签进行语义扩充，得到扩充后的标签语义集。

(3) 微博数据处理。利用 ICTCLAS②分词系统对微博进行分词；同时，在分词时将用户标签和标签扩充词作为分词词典导入分词系统中以提高分词性能。

(4) 用户标签与微博内容的匹配。以用户标签集作为匹配词典，标签集与用户的微博内容进行匹配。

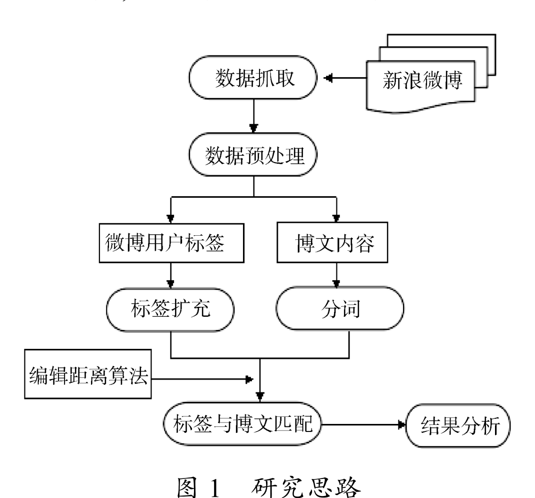


图 1　研究思路

3.2　数据处理

(1) 微博数据预处理

在数据准备过程中，过滤标签数为 0 的用户，最后得到 760 位用户共 703 635 条微博。并将连续转发和评论的微博看做是一条完整的微博来处理。

①利用微博 API 接口获取用户自定义的标签，此次获取 835 个用户，去掉标签量为 0 的用户，得到 760 位用户共 4 689 个标签，人均标签约 6 个。

②利用梁斌研发的词库 API 对用户的每个标签进行语义扩充。表 1 所示为“自然语言处理”的部分扩充结果，“人工智能”为扩充词语，“0.184195”表示词语相关程度。

③将抓取的标签和标签扩充词进行处理，组成标签集，为分词和博文匹配做准备。

①http://cikuapi.com.

②http://ictclas.nlpir.org/.

表 1 微博用户原标签及标签扩充结果表

用户标签	扩充结果
自然语言处理	人工智能, 0.184195 机器翻译, 0.148667 自然语言理解, 0.138319 中文信息处理, 0.119663 语音识别, 0.108192 计算机科学, 0.108003 模式识别, 0.105456 数据挖掘, 0.097883 智能, 0.092609 中文分词, 0.089793 表达式, 0.057491 哈尔滨工业大学, 0.057463

(2) 微博博文处理

在上述数据基础上，对用户博文进行分词，这也是文本处理的基础。目前，中文分词有很多算法和工具，本文通过 ICTCLAS 分词系统对博文分词，得到结果中的词都带有词性标记，比如名词/n、动词/v、形容词/a。而用户标签基本以名词为主，例如某位微博用户给自己打的标签：情感分析、自然语言处理、数据挖掘、文本分类、模式识别、乒乓球、博士、扬州、南京、北京。ICTCLAS 分词系统允许用户导入词典，本研究将上面得到的标签集作为分词词典导入到分词系统中。系统词典的格式是按行排列，且词语后带有词性。例如，博士 n，词语和词性标识之间是一个 Tab 键。在处理数据时，将所有的标签词都标识为 tag。加入标签词典和未加入词典的前后处理效果对比，如表 2 所示：

表 2 微博内容分词结果对比表

比较项	结果
未加入标签词典	《/wkz MIT/x 自然语言/un 处理/un 讲座/un 》/wky 这会/un 不会/un 成为/un 一个/mq 新的/un 高等/un 批判/un 法/n 的/ude1 发端/un 呢/y
加入标签词典	《/wkz MIT/n 自然语言处理/tag 讲座/n 》/wky 这/rzv 会/v 不/d 会/v 成为/v 一个/mq 新/a 的/ude1 高等/b 批判/vn 法/n 的/ude1 发端/n 呢/y

可以看出，加入标签集后标签集中出现的短语会被切成一个整体，未加入标签词典的分词结果将“自然语言处理”切分成“自然语言”和“处理”；加入标签集后“自然语言处理”则被切分成一个短语。

3.3 标签–博文关联匹配

对分词后的博文进行处理，在微博数据准备中共有 703 635 条微博文本和 23 487 条用户标签，对这些数据按照用户 ID 进行汇总，得到每个用户的微博集合和标签集合，然后进行博文匹配。

以扩充后的用户标签集作为匹配词典，将标签集与用户的微博内容进行匹配，以实现标签词语与微博文本语义匹配。计算词语之间语义相似度的方法有很多，比如基于语料库[11]、基于词典[12]、基于网络或本体[13-14]以及基于编辑距离[15]等。为了简化处理，本研究利用词库 API 进行标签词的语义扩充，同时将扩充词引入到分词词典中提高分词精度来保障前期处理的效果，从而在后期词语相似匹配时选择易于使用的编辑距离算法来进行用户标签集与用户微博内容的匹配。编辑距离(Edit Distance)[15]，又称 Levenshtein 距离，是指两个字串之间，由一个转成另一个所需的最少编辑操作次数。Sim(Tag(u), Word(v))：表示标签 u 和微博词语 v 之间的相似度。令 Tag(u)，Word(v)的编辑距离为 Distance(Tag(u)，Word(v))，length(x)表示 x 的长度，则 Tag(u), Word(v)之间的相似度计算公式如下：

$$\mathrm{Sim}(\mathrm{Tag}(u),\mathrm{Word}(v)) = 1 - \frac{\mathrm{Distance}(\mathrm{Tag}(u),\ \mathrm{Word}(v))}{\mathrm{Max}(\mathrm{length}(\mathrm{Tag}(u)),\ \mathrm{length}(\mathrm{Word}(v)))} \tag{1}$$

由于编辑距离反映两个字符串的绝对差异，且受词语长度的影响，因此在进行数据处理时只对标签长度大于 3 个字符的词语进行相似度计算，长度小于等于 3 的标签词语则进行字面匹配。通过反复实验验证发现当相似度取大于 0.5 的时候的匹配结果最为理想。

以微博 ID 为 1065269410 的用户为例来说明标签与微博内容的匹配结果。结果如表 3 所示：

表 3 标签与博文关联分析示例

用户标签	原博文	匹配结果
情感分析 自然语言处理 数据挖掘 文本分类 乒乓球 模式识别 博士 扬州 南京 北京	//@张家俊 MT：赞//@KJ 音乐人生_王亮_自动化所：转发微博【谭铁牛当选英国皇家工程院外籍院士】9 月 15 日，在英国皇家工程院年会上，中国科学院副秘书长谭铁牛当选英国皇家工程院外籍院士。谭铁牛是中国科学院院士，计算机视觉与模式识别领域专家。	计算机视觉 模式识别 专家 院士 英国 中国科学院

在匹配结果中，“模式识别”是用户给自己打的标签，“计算机视觉”是“模式识别”的扩充词。“专家”、“院士”、“英国”、“中国科学院”都是标签词语“博士”的扩充词。从匹配结果“计算机视觉 模式识别 专家 院士 英国 中国科学院”中可以看到“模式识别”是这类用户自己打的标签，以及“计算机视觉 专家 院士 英国 中国科学院”是这类标签扩充词。

4　实验结果分析

4.1　专业领域用户添加标签行为的分析

通过对用户添加个人标签行为的统计，得到如图 2 所示结果。835 位用户中有 760 位用户至少添加了一个标签，只有 75 位用户没有为自己添加任何标签。这与邢千里等[10]以普通用户为研究对象得出的结果有较大差别，其得出在普通用户标签数量分布中，有 59.4%的用户没有为自己添加任何标签。同时，本研究还发现在添加标签的用户里，有 572 位用户的标签数量都在 5 个以上。因此，可以认为专业领域用户更乐意为自己添加标签并且也愿意为自己添加尽可能多的标签来获得同行关注。

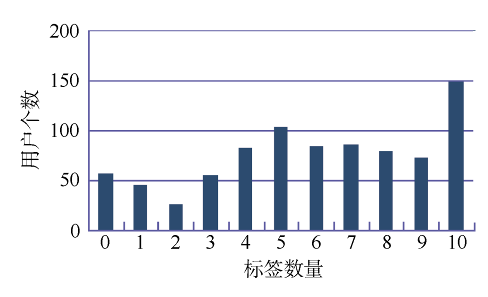


图 2　用户标签数量分布

同时，通过对“自然语言处理”领域的用户添加的标签内容进行分析，除去抓取数据时选取的关键词“自然语言处理”和“中文信息处理”(其中“自然语言处理”出现的频次为 622，“中文信息处理”的频次为 35)，得到用户标签使用频次前 20 的标签，如表 4 所示:

表 4　用户标签使用频次

标签	使用频次	标签	使用频次
机器学习	260	美食	43
数据挖掘	190	80 后	39
信息检索	111	推荐系统	39
IT 数码	93	机器翻译	38
互联网	70	文本挖掘	37
搜索引擎	68	IT	33
NLP	64	计算机	33
旅游	59	大数据	32
计算语言学	50	电影	32
人工智能	48	音乐	29

可以看出，这些高频词多为和“自然语言处理”领域相关的专业性术语，而像大众性的标签描述只出现“旅游”、“美食”、“80 后”、“电影”、“音乐”这 5 个。这些标签的出现是因为此类标签最容易作为系统推荐标签出现，不用手动输入，而且这类标签对于用户具有普适性。与邢千里等[10]研究中普通用户所使用的热门标签进行对比发现，专业领域的用户更倾向于用专业性较强的标签词语来描述自己的专业领域，而不是直接在标签推荐页面不加思考地选择热门推荐标签。

4.2　专业领域微博用户标签与博文相关度度量

鉴于微博文本的特殊性，现有的主题模型并不能很好地分析微博的内容，本文没有采用关键词提取的方法进行标签与微博内容的匹配，而是直接将标签与微博内容进行语义匹配。对博文匹配结果进行统计，结果如图 3 所示。其中微博和标签匹配率=与用户标签关联的微博数/用户所发布的微博总数。

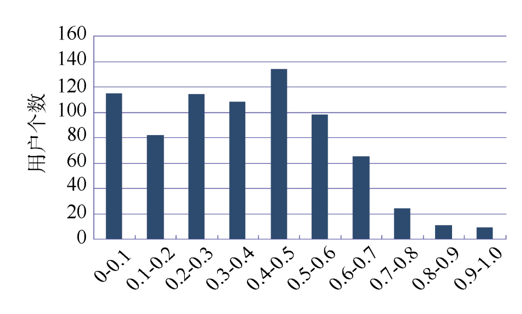


图 3　博文匹配结果统计图

从图 3 可以看出，用户微博和标签匹配率主要集中在 70%以下，只有少数用户的匹配率达到 70%以上。760 名用户中，有 341 位用户的微博和标签匹配率达到 40%以上。由此可知，用户标签与微博内容有一定的关联程度。

去除标签和微博数的影响，选取高匹配率区间(匹配率大于 0.7)的用户微博内容和低匹配率区间(匹配率小于 0.1)的用户微博内容进行词频统计。表 5 为词语出现频次前 20 的词语，以是否与“自然语言处理”领域相关来判断词语的专业性，发现高匹配率区间 Top20 中词语的专业相关度要高于低匹配率区间。同时，可以看出高匹配率区间的词语频次要远高于低匹配率区间，根据博文中词语频次统计按降序排列，分别取出现次数 Top100, Top500, Top1000 的词语进行观测，均呈现出以上规律。同一词语出现的频次一定程度上反映的是微博内容之间的相似度，因此得出高匹配率区间用户所发微博内容间的相似性要高于低匹配率区间。

表 5　不同匹配区间微博内容词频 Top20 的分布

高匹配率区间				低匹配率区间			
词语	频次	词语	频次	词语	频次	词语	频次
技术	4 928	搜索	2 893	喜欢	1 811	孩子	1 280
数据	4 767	时间	2 892	时间	1 758	技术	1 193
公司	4 311	语言	2 800	美国	1 663	学习	1 069
课程	4 199	文章	2 788	世界	1 547	数据	1 043
用户	4 166	百度	2 745	手机	1 524	苹果	1 025
同学	3 345	应用	2 714	老师	1 485	同学	1 014
老师	3 286	信息	2 706	公司	1 417	学生	1 006
学习	3 089	翻译	2 638	生活	1 384	应用	910
机器学习	3 052	大数据	2 446	用户	1 323	小时	868
算法	2 999	NLP	2 405	北京	1 323	百度	863

表 6 为词语频次统计表中 Top20, Top100, Top500 和 Top1000 中的专业性词语的数量, 从两个区间专业性词汇比例分布情况可以看到, 高匹配率区间用户专业性词汇的分布呈现逐渐递减的趋势, 也就是说该区间的词语频次分布越靠前专业性词汇越多, 而低匹配率区间总体呈现均衡状态, 说明该区间词语分布较为离散。通过对这些区间用户所发微博内容的观察, 发现高匹配率区间的用户所发微博内容多为科研信息、行业见闻或者资讯分享; 而低匹配率区间的用户所发微博内容多为生活状态、感情抒发, 微博内容丰富多样。这就导致高匹配率区间的专业性词汇占比和词语频次都高于低于匹配率区间。

表 6　不同比例区间微博内容词频分布

比较项	Top20		Top100		Top500		Top1000	
	高	低	高	低	高	低	高	低
专业词汇量	9	2	41	18	175	85	296	155
专业词占比	45%	10%	41%	18%	35%	17%	29.6%	15.5%

(1) 原因分析

用户标签数和所发微博数是影响匹配结果的直接因素, 对用户的标签数、所发微博数进行观察分析发现: 匹配率为 0 的 51 位用户里, 有 27 位用户的只有 1 个标签, 其余用户所发微博数均小于 40; 将标签数小于 5, 微博数小于 100 的用户去除, 得到用户粉丝数在不同匹配区间上的分布情况, 如图 4 所示, 虽然整体上没有呈现逐渐增长的趋势, 但可以看出匹配率大于 0.6 的用户粉丝平均数要明显高于小于 0.6 的。粉丝数代表用户的影响力, 也可以反映出用户的活跃程度, 由此可知, 影响力大的用户微博和标签的关联程度要高于影响力小的用户。

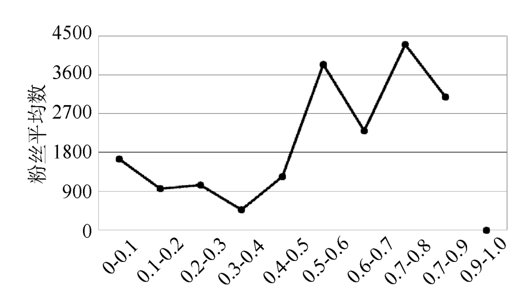


图 4　用户粉丝数与匹配率关系图

选取图 4 中匹配率大于 0.6 和小于 0.2 的用户标签进行分析, 按照标签词语的专业性程度分为专业性词汇和非专业性词汇两类。通过数据调查发现匹配率大于 0.6 的专业性标签词汇占比为 81%, 匹配率小于 0.2 的专业性标签词汇占比为 65%。同时发现匹配率小于 0.2 的非专业词汇中有非常多的网络用语, 而在匹配率大于 0.6 的非专业词汇中只发现了一个。例如, 在描述程序员这个职业时出现了: IT 女精英、挨踢女民工、码农小伙伴、软件攻城师等词汇。不同匹配区间标签词汇对比如表 7 所示:

表 7　不同匹配区间的标签词汇对比

类别	匹配率大于 0.6	匹配率小于 0.2
饮食类	餐饮　美食	喜欢在家吃饭　吃货　食色性也　吃吃吃吃吃　美食　爱咖啡如命　美食爱好者
追星类	潮人　科比	业余菲迷　苏轼粉　五月天　fans　巴萨球迷
视频类	电影　美剧	美剧重症患者　看电影　动漫控

通过对比发现, 匹配率大于 0.6 的用户标签词汇除了一个“天然萌”其余全部为传统汉语词汇。而匹配率小于 0.2 的 168 个非专业词汇中有 70 个为网络用语或者描述性的短句。由此可知, 专业性词汇的占比再结合非专业性词汇的特性在一定程度上能够反映用户的专业性程度, 专业性程度高的用户微博和标签的关联程度要高于专业性程度低的用户。

(2) 不同匹配标签集的匹配结果分析

将用户原标签以及扩充后的标签集分别与博文进行匹配, 结果如图 5 所示。与扩充后的标签集的匹配比例相比, 原标签与博文的匹配率较低, 主要集中在 10%以下。可以看出, 原标签和用户所发微博有一定关

联，只是相关联的微博条数较少。主要由于用户标签最多只有 10 个词语，表达能力非常有限，这也从侧面说明对标签进行语义扩充的合理性。

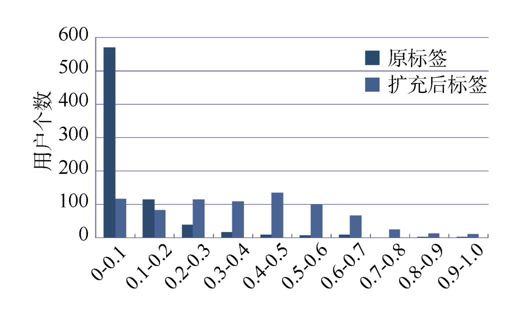


图 5　不同标签集的匹配结果对比

以上结果表明，用户原标签与博文有一定的相关度，但是对标签进行扩充后的相关度要远高于原标签的相关度。本文通过编辑距离算法和扩充标签词集实现词语与博文内容的匹配，不是将原标签与博文进行直接字面匹配，而是实现了标签与博文的语义匹配，得到基于语义的匹配结果要优于基于原标签的匹配结果，使得结果更加合理。

5　结　语

本文以新浪微博中学术型用户微博为例，采集微博用户的用户标签和微博内容数据，利用这些数据对自然语言处理领域微博用户添加标签的行为特点和标签内容进行分析，结果表明专业领域的用户更乐意为自己添加尽可能多的标签，也更倾向于使用专业性较强的词语作为自己的标签；利用这些数据进行用户标签与微博内容关联度统计分析表明：在新浪微博平台上，学术领域微博用户标签和用户微博内容具有一定的相关度。同时，除了微博数和标签数外，用户的影响力、专业性程度都会对用户标签和用户微博内容相关度产生影响。

通过本文的研究，建议一般微博用户能够像学术领域的用户一样重视自己定义的标签，避免随意地给自己打上热门标签，或者不愿意花费时间为自己打标签。同时，微博标签推荐系统可考虑将用户微博内容作为标签推荐的重要数据来源，为用户提供更有针对性的个性化标签。反之亦然，在对微博内容进行主题分析时，可以借助于微博用户标签优化博文主题发现和分析结果。

在未来的研究中，将进一步研究用户标签在微博主题推荐、微博信息检索、用户建模等方面的应用，同时扩大研究对象的领域和社会化应用平台。

参考文献：

[1] Al-Khalifa H S, Davis H C. Folksonomies Versus Automatic Keyword Extraction: An Empirical Study [J]. IADIS International Journal on Computer Science and Information Systems, 2006, 1(2): 132-143.

[2] Rolla P J. User Tags Versus Subject Headings [J]. Library Resources & Technical Services, 2011, 53(3): 174-184.

[3] Thomas M, Caudle D M, Schmitz C M. To Tag or not to Tag? [J]. Library Hi Tech, 2009, 27(3): 411-434.

[4] Lu C, Park J R, Hu X. User Tags Versus Expert-assigned Subject Terms: A Comparison of LibraryThing Tags and Library of Congress Subject Headings [J]. Journal of Information Science, 2010, 36(6): 763-779.

[5] 潘婵，冯利飞，丁婉莹，等. 基于标签–关键词的用户行为分析[J]. 情报杂志，2010，29(3)：139-142. (Pan Chan, Feng Lifei, Ding Wanying. Tag and Keyword-Based Analysis of Users' Behavior [J]. Journal of Intelligence, 2010, 29(3): 139-142.)

[6] Kipp M E I. Tagging of Biomedical Articles on CiteULike: A Comparison of User, Author and Professional Indexing [J]. Knowledge Organization, 2011, 38(3): 245-261.

[7] Lee D H, Schleyer T. Social Tagging is no Substitute for Controlled Indexing: A Comparison of Medical Subject Headings and CiteULike Tags Assigned to 231, 388 Papers [J]. Journal of the American Society for Information Science and Technology, 2012, 63(9): 1747-1757.

[8] 黄红霞，章成志. 中文微博用户标签的调查分析——以新浪微博为例[J]. 现代图书情报技术，2012(10)：49-54. (Huang Hongxia, Zhang Chengzhi. Investigation and Analysis of Chinese Microblog User Tags——Using Sina Weibo as Example [J]. New Technology of Library and Information Service, 2012(10): 49-54.)

[9] 章成志，何陆琳，丁培红. 不同领域的用户标签主题表达能力差异研究——以中文微博为例[J]. 情报理论与实践，2013，36(4)：68-71. (Zhang Chengzhi, He Lulin, Ding Peihong. Difference of Subject Expression Function of User Tags in Different Domains——Using Chinese Microblogging as Example [J]. Information Studies: Theory & Application, 2013, 36(4): 68-71.)

[10] 邢千里, 刘列, 刘奕群, 等. 微博中用户标签的研究[J]. 软件学报, 2015, 26(7): 1626-1637. (Xing Qianli, Liu Lie, Liu Yiqun, et al. Study on User Tags in Weibo [J]. Journal of Software, 2015, 26(7): 1626-1637.)

[11] Baeza-Yates R, Ribeiro-Neto B. Modern Information Retrieval [M]. New York: ACM Press, 1999.

[12] Kozima H, Furugori T. Similarity Between Words Computed by Spreading Activation on an English Dictionary [C]. In: Proceedings of the 6th Conference on European Chapter of the Association for Computational Linguistics. Association for Computational Linguistics, 1993: 232-239.

[13] 江敏, 肖诗斌, 王弘蔚, 等. 一种改进的基于《知网》的词语语义相似度计算[J]. 中文信息学报, 2008, 22(5): 84-89. (Jiang Min, Xiao Shibin, Wang Hongwei, et al. An Improved Word Similarity Computing Method Based on HowNet [J]. Journal of Chinese Information Processing, 2008, 22(5): 84-89.)

[14] Budanitsky A, Hirst G. Semantic Distance in WordNet: An Experimental, Application-oriented Evaluation of Five Measures [C]. In: Proceedings of the Workshop on WordNet and Other Lexical Resources, the 2nd Meeting of the North American Chapter of the Association for Computational Linguistics, Pittsburgh. 2001.

[15] Levenshtein V I. Binary Codes Capable of Correcting Deletions, Insertions, and Reversals [J]. Soviet Physics Doklady, 1966, 10(8): 707-710.

作者贡献声明:

薛春香: 提出研究思路, 设计研究方案, 论文最终版本修订;
章成志: 数据采集, 研究思路讨论;
朱玲: 数据分析和处理, 起草论文;
傅柱: 研究思路讨论, 论文修改建议。

利益冲突声明:

所有作者声明不存在利益冲突关系。

支撑数据:

支撑数据由作者自存储, E-mail: xuechunxiang@njust.edu.cn。

[1] 朱玲, 薛春香, 章成志, 傅柱. tagNum.xlsx. 用户标签数量分布.

[2] 朱玲, 薛春香, 章成志, 傅柱. tagFreq.xls. 用户标签使用频次.

[3] 朱玲, 薛春香, 章成志, 傅柱. wordFreq.xls. 不同匹配区间微博内容词频分布.

[4] 朱玲, 薛春香, 章成志, 傅柱. funsNum.xlsx. 用户粉丝数在不同匹配区间上的分布.

[5] 朱玲, 薛春香, 章成志, 傅柱. tagComp.xlsx. 不同匹配区间的标签词汇.

[6] 朱玲, 薛春香, 章成志, 傅柱. matchResult.xls. 原标签以及扩充后的标签集与博文匹配结果.

收稿日期: 2015-09-14
收修改稿日期: 2015-10-24

User Tags and Microblog Posts: Case Study of Sina Weibo

Zhu Ling Xue Chunxiang Zhang Chengzhi Fu Zhu
(School of Economics and Management, Nanjing University of Science and Technology, Nanjing 210094, China)

Abstract: [**Objective**] This study aims to explore the relationship between the user tags and microblog post topics, with the purpose of improving subject identification and automatic tag recommendation services. [**Methods**] We first used crawlers to retrieve user profiles and posts in the field of “natural language processing” from the Sina Weibo. Second, extracted words from the posts and semantically extended user tags. Finally, matched the tags and posts by the edit distance algorithm. [**Results**] There was correlation between user tags and posts in natural language processing field. [**Limitations**] We only studied one academic field and the Sina Weibo, more research is needed in the future to generalize the results. [**Conclusions**] The tag recommendation system can use microblog posts as an important source to provide more personalized services, which in turn will improve the microblog content analysis.

Keywords: Subject analysis of posts User tags Correlation measure Subject indexing User modeling